Paige Rattenberry

AI/ML systems engineer

paigerattenberry@gmail.com | linkedin.com/in/paige-rattenberry | github.com/PaigeRattenberry | paigerattenberry.vercel.app | Burnaby, BC, Canada

AI/ML Systems Engineer building the infrastructure that lets AI agents act reliably — MCP servers, agent loops with citation guardrails, LLM grounding and evaluation enforced as CI merge gates, and retrieval spanning BM25F search and retrieval-augmented generation (RAG). Full-stack across Python/FastAPI and TypeScript/Next.js; primary engineer across 300+ pull requests on an internal knowledge platform, with a background in safety-critical medical AI.

TECHNICAL WORK EXPERIENCE

Software Engineer – Advanced Research Engineering — May 2025 – Present
Hammerspace, Burnaby, BC (Remote)

- Built grounding and evaluation infrastructure to make LLM output verifiable — deterministic checks and LLM-judge validation running as CI merge gates, a scheduled audit job opening its own remediation pull requests, and server-side access-control filtering with human review gating AI-authored content.
- Designed and built an internal knowledge platform end-to-end (Python/FastAPI, Next.js/TypeScript) — ingestion pipelines across multiple document and collaboration sources, an LLM synthesis workflow where every published claim traces back to an immutable source at fragment granularity, and a review, search & graph-exploration web app, front end included; demoed the platform to leadership; primary engineer across 300+ pull requests.
- Built the retrieval and conversational layer — an MCP server under access-control filters, BM25F search with query expansion, an offline retrieval evaluation harness over golden queries, and a streaming (SSE) chat service with an agent loop and citation guardrails on model output.
- Contributed to a multi-agent infrastructure-automation platform on the OpenAI Agents SDK (containerized agents, LLM-based routing between specialized agents, MCP tool integrations, guardrails and permissioning) and wrote concurrent LLM API workloads at volume; fine-tuned a 7B code-generation model (LoRA, CodeLlama-7B-Instruct, in PyTorch with Hugging Face PEFT), building its instruction dataset and a ~95-query evaluation suite with accuracy and latency logging, containerized for remote GPU inference.

Software Engineer – M365 Core Substrate SHIELD Security — Sept 2024 – Feb 2025
Microsoft, Redmond, WA, USA

- Engineered a security-focused anomaly detection & threat intelligence system using sequential log data, applying temporal pattern analysis to proactively detect & mitigate emerging cyber threats.
- Partnered with security & infrastructure teams to design and deploy automated, scalable models across live telemetry, translating asynchronous sequence data into actionable risk profiles; triaged and mitigated live security incidents on production services.

Software Engineer – Azure Cloud Storage Performance Infrastructure & Analytics — Nov 2022 – Sept 2024
Microsoft, Redmond, WA, USA

- Designed and developed AI-powered system diagnostics, optimizing large-scale, cloud-based retrieval and anomaly detection models for distributed AI systems.
- Built telemetry analytics pipelines in SQL & KQL and extended distributed C# monitoring across petabyte-scale Azure storage, using statistical analysis to find latency bottlenecks and cut storage access latency.

Machine Learning Engineering Co-op – Computer Vision — Sept 2021 – May 2022
Stryker R&D, Burnaby, BC

- Designed, trained, and evaluated CNN-based computer vision models in PyTorch for surgical video analysis, toward intra-operative complication prediction in complex, high-risk procedures.
- Applied and evaluated explainable AI (XAI) techniques in PyTorch — including saliency mapping, feature attribution, and counterfactual reasoning — to improve model transparency and support surgeon decision-making.
- Researched multimodal computer vision methods which integrate surgical video with structured clinical knowledge, to improve model robustness and interpretability in safety-critical, domain-specific settings.

Software Engineering Intern & Explore Intern – Microsoft Azure Cloud Storage — May – Jul 2021 & May – Aug 2019
Microsoft, Virtual & Redmond, WA

- Built low-overhead C++ latency-sampling tools and Python log-processing automation to tune Azure Cloud Storage performance (2021), and a modular performance-testing framework benchmarking the Azure .NET SDK (2019).

Product Management Co-op — Jan – Apr 2019
Samsung R&D Canada, Vancouver, BC

- Researched AI/ML paradigms and emerging frameworks to inform Samsung's enterprise product strategy and roadmap planning.

COMPUTER VISION, AI RESEARCH & ENGINEERING PROJECTS

StimMap3D – Illustrative model — not for clinical use — Jun – Aug 2026
Interactive 3D model of TMS coil placement and the cortical field it induces

- Independently designed and built a browser-based, interactive 3D visualization tool modeling TMS (transcranial magnetic stimulation) coil placement and induced cortical electric fields, with a live-recomputing heatmap driven by coil drag, orbit, and zoom interaction — including focality iso-contour bands, field-direction glyphs, and depth-dose readouts across configurable targeting presets.
- Built supporting UX for scientific communication (guided onboarding tour, shareable deep-links, PNG export, print-ready clinician handouts, and a “Methods & Limitations” page with a self-check quiz), framed as an illustrative, non-clinical model.

Evaluation of Feature Attribution Methods in Interpretable Machine Learning — Apr – Aug 2022
SFU Engineering Science | Supervisor: Dr. Ivan V. Bajić (SFU Multimedia Lab)

- Built a machine learning framework in PyTorch to evaluate the interpretability of deep neural networks (DNNs) for medical image classification in safety-critical applications, training and benchmarking ResNet-34 and VGG-16 models.
- Applied Zhou et al.'s induced-ground-truth design (AAAI 2022) to chest X-rays: watermarked them, then used IoU and paired t-tests to measure how precisely five CAM methods localized the watermark; found FullGrad the most robust and precise, improved further by thresholding.

Adolescent Spinal Curvature Brace Optimization
Jan – Aug 2022

SFU Engineering Capstone; Best Overall Project, ICAMES 2022

- Originated the project and formed the six-person team that built a wearable pressure-sensing system (force-sensor array, noise-filtering firmware, real-time pressure heatmap on a CAD torso model) to guide orthopedic brace design, following medical-device software standards (IEC 62304).
- Won Best Overall Project at ICAMES 2022 (Boğaziçi University, Istanbul), an international engineering student competition; proposed an ML model trained on the system's pressure data to predict curve progression as the next phase.

This site, built with Claude Code under CI gates
Sep 2026 – Present

Personal project

- Directed Claude Code through every session as a reviewed pull request, with Codex (GPT-6) for one review and its follow-up, making the decisions and merging every PR, and held the agents' work to CI gates: a claims gate on unsourced metrics, a publication gate on private paths and terms, and accessibility and performance limits.

Interpretability and Explainability in Generative AI
Apr 2024

Single-author literature review

- Reviewed what interpretability and explainability research offers against generative AI's failure modes (hallucination, misgeneralization, bias, prompt injection, deepfakes), drawing on 14 papers, from model-graded resilience scoring and causal-inference benchmarking to saliency-aware counterfactuals and rule extraction from reinforcement-learning agents. Concluded that no method was yet widely deployed, and that standardized interpretability metrics weighted by application domain are essential to deploying these systems with confidence.

Clinical CoPilot
Sep 2023

Microsoft Global AI Hackathon

- Built a retrieval-augmented generation (RAG) pipeline for a clinical LLM using Azure OpenAI and Semantic Kernel; ranked 16th/1,253 teams with an Executive Challenge honourable mention at Microsoft's Global AI Hackathon.

Computer Vision, 3D Reconstruction & Segmentation
Sept – Dec 2021

SFU CMPT 412, graded A+: classification, detection, segmentation and multi-view 3D reconstruction

- Conducted image classification, object detection, and semantic/instance segmentation using PyTorch & OpenCV; ranked 2nd/84 in a Kaggle competition by optimizing CNNs for segmentation.

AI LEADERSHIP & RESEARCH CONTRIBUTIONS

Microsoft AI Expert Panel Q&A – Organizer & Host
Jul 2024

Microsoft

- Organized & hosted Microsoft's AI Expert Panel on LLM scaling, retrieval architectures, explainability, and responsible AI.

MLADS Co-Area Chair & Proposal Reviewer
Jun 2024

Microsoft Machine Learning, AI & Data Science Conference

- As co-area chair & proposal reviewer for Microsoft's MLADS AI/ML conference, evaluated and selected talks and posters in forecasting, anomaly detection, and semi-supervised learning.

Microsoft Aspire Leadership Council Member | Azure Core Early-in-Career Pillar Lead
Jul 2023 – Feb 2025

Microsoft

- Served on Microsoft's Aspire Leadership Council, leading Azure Core's early-in-career pillar and mentoring early-career engineers.

Researcher – Rostrum Medical Innovations & HealthTech Connex
Aug 2020 – Apr 2021

Rostrum Medical Innovations Inc.; HealthTech Connex Inc. (BrainNet)

- Researched capnography classification and EEG-based disease detection with interpretable multimodal deep learning for explainable medical AI.

Medical Imaging Research Assistant
May – Aug 2018

SFU Functional & Anatomical Imaging & Shape Analysis (FAISAL) Lab

- Processed and analyzed CT scan data, applying segmentation correction techniques relevant to computer vision.

TECHNICAL SKILLS

Full-stack & infrastructure: Python, FastAPI, TypeScript, Next.js / React, C#, C++, SQL, KQL, REST APIs, Docker & containerization, CI/CD, Azure, Azure DevOps, Access control & permissions, Android (Java)

Agentic AI & LLM systems: MCP, Agent loops & tool calling, OpenAI Agents SDK, Semantic Kernel, LangChain, Azure OpenAI, Gemini / Vertex AI, BM25F search & query expansion, RAG, LLM evaluation harnesses, CI-gated validation, Grounding & claim verification, LoRA fine-tuning, MLOps, Generative AI

ML, computer vision & medical imaging: PyTorch, TensorFlow, OpenCV, Deep learning, CNNs, Computer vision, Image classification, Segmentation, Medical imaging, Anomaly detection, Time-series analysis, Interpretability (XAI), Biomedical signals (EEG, capnography), Firmware & sensors

Tooling: Claude Code, Codex, Git, GitHub Actions

EDUCATION, CERTIFICATIONS & AI SPECIALIZATIONS

Bachelor of Applied Science – Computer Engineering (Honours with Distinction)
Sept 2017 – Aug 2022

Simon Fraser University, Burnaby, BC, Canada

- GPA: 3.89 (top 5%, Faculty of Applied Sciences) | President's & Dean's Honour Rolls | Dr. Abe Unrau Memorial Co-op Award
- Valedictorian (45:27) – selected to deliver the graduand address at the October 2022 convocation.
- Co-creator and speaker, SFU's “From Academia to Industry” podcast on internship and career guidance (Nov – Dec 2022).

Certifications

- GIAC/SANS.edu: GIAC Machine Learning Engineer (GMLE), Dec 2024 – Feb 2025 — Scored 99%, with the top rating in 13 of 14 domains, earning a GIAC Advisory Board invitation; via SANS SEC595, applying MLOps practices to malware, anomaly and fraud-detection models.
- SANS Technology Institute: SEC595: Applied Data Science & AI/Machine Learning for Cybersecurity Professionals, Completed 9 Jan 2025 — The course behind the GMLE certification, completed for 36 CPE credits.
- Coursera: Multi-Modal AI specialization, 2021 – 2022; Deep Learning specialization, Dec 2021; Natural Language Processing specialization, Jan 2022; AI for Medicine specialization, Jan 2022
- Google / Kaggle: Google 5-Day Gen AI Intensive Course Certificate, Apr 2025